符号理論とは

情報数理学特別講義I 第1回

2025-11-04

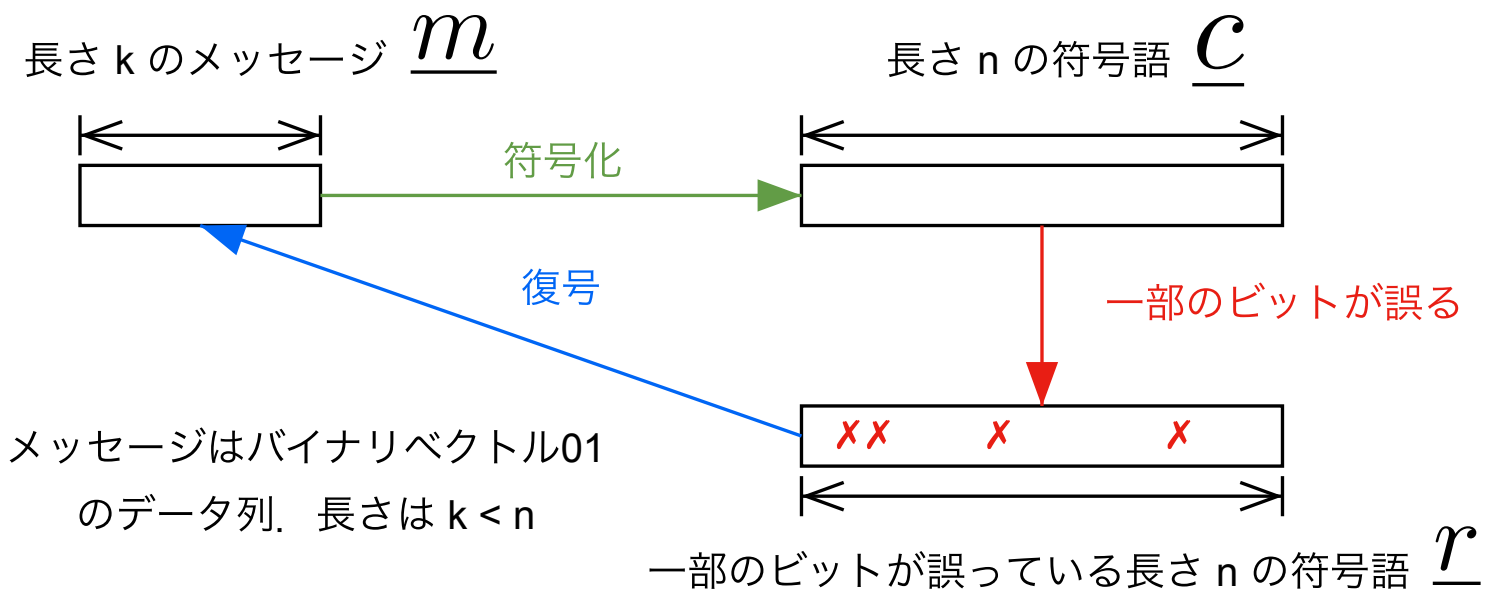
符号理論とは

情報理論の中の一分野で，通信路符号化技術に関する理論.
符号理論で扱う**符号**には，誤りを**検出**するための**誤り検出符号**や，
誤りを訂正するための**誤り訂正符号**がある.

誤りの定義にも様々ある．例えば消失誤り，反転誤り，削除誤りなど．それぞれの誤りに対する符号が研究されている．例えば消失訂正符号，誤り訂正符号，削除誤り訂正符号.

誤り訂正符号は1950年代後半から**代数**を駆使して理論が発展.
1990年からは**確率推論**にもとづく符号理論も発展.
非常に大きな理論体系で，現在も理論の深耕，拡大が続いている.
符号理論の歴史については以下の文献が詳しい
今井秀樹, “誤り訂正符号化技術の歴史,” 映像情報メディア学会誌 70 (4), pp.562-566. 2016.

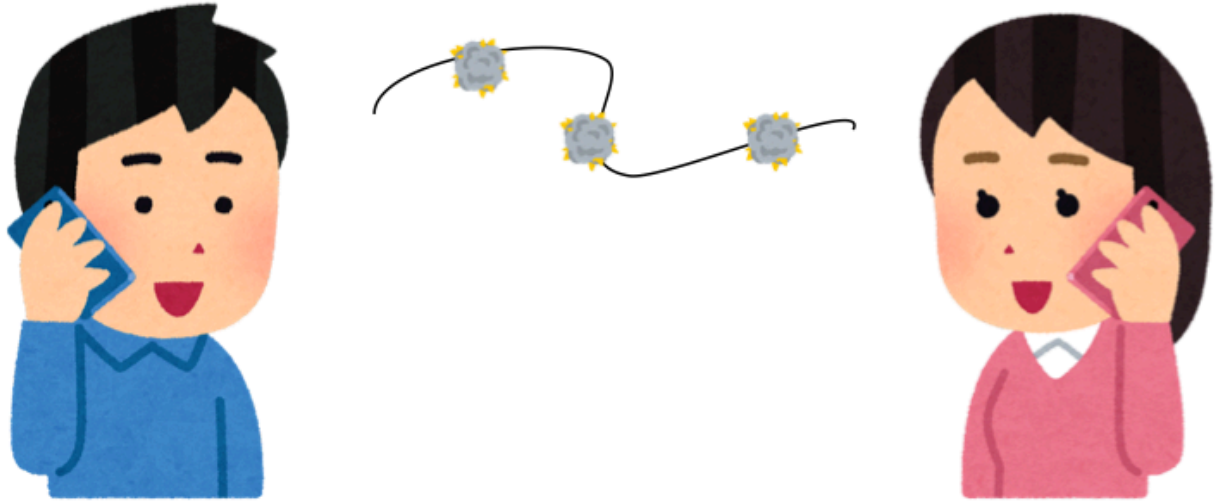
符号理論の問題設定



符号化 によりメッセージの長さを伸ばす, $n > k$．冗長な情報を付加する．
誤りを含む $\tilde{c}$ から，もとのメッセージ x を推測する（**復号**）．
工学的には，少ない冗長（$n - k$が小さい）で沢山の誤りへの耐性を持ち，
容易に符号化＆復号が可能な誤り訂正符号を実現したい．

符号は様々な情報システムで用いられている

安定した通信の実現



メディア故障からの保護

符号とは

符号語(codeword)と呼ばれるビット列 $001011100\cdots$ の集合

- 一般にはビットに限らず，任意の記号列の集合.
- この講義では主に系列の要素が 0 と 1 の2つの記号からなるビット列の集合を扱う.

例：長さ3の単一パリティ検査符号(Single Parity Check Code) $\mathcal{C} = \{000, 101, 011, 110\}$
※長さ 3 とは，符号語の長さが 3 であることを意味する.

符号 $\mathcal{C}$ には4つの符号語が含まれる.
4つの符号語で表現できるメッセージは2ビット($\log_2 4 = 2$)なので,
この符号のメッセージ長 k は2ビットである.

長さ3の単一パリティ検査符号

$\mathcal{C} = \{000, 101, 011, 110\}$, 符号語の数は 4.

単一パリティ検査符号の符号語は長さ2のメッセージ $\{00, 10, 01, 11\}$ に対応している.

メッセージ	符号語
00	000
10	101
01	011
11	110

このようにメッセージを符号語に変換することを **符号化（encode）** という.

長さ3の単一パリティ検査符号

$\mathcal{C} = \{000, 101, 011, 110\}.$

問題：単一パリティ検査符号の符号語が持つ特徴はなんでしょう？

単一パリティ検査符号 (Single Parity Check Code)

答え：

例：長さ3の単一パリティ検査符号 $\mathcal{C} = \{000, 101, 011, 110\}$

単一パリティ検査符号の符号語に含まれる 1 の数（**重み**と言う）が偶数のため，**奇数ビットの誤りが検出できる.**

- ここでの誤りは $0 \to 1$ や $1 \to 0$ のビットの反転誤りを指す
  （特に断らない限り，本講義での誤りはビット反転誤り）.

単一パリティ検査符号の誤り検出能力を確認

例：長さ3の単一パリティ検査符号 $\mathcal{C} = \{000, 101, 011, 110\}$

符号語 101 に対して1ビット誤りが発生した場合

- 左から1ビット目が誤ると $\color{red}{0}01$. 重みが1となり奇数になることから誤りを検出.
- 左から2ビット目が誤ると $1\color{red}{1}1$. 重みが3となり奇数になることから誤りを検出.
- 左から3ビット目が誤ると $10\color{red}{0}$. 重みが1となり奇数になることから誤りを検出.

2ビット誤るとどうなるか？

- 左から1,2ビット目が誤ると $\color{red}{01}1$. となり，符号語と一致し誤りを検出できない.

符号の効率：符号化率

符号語に占めるメッセージの割合を符号化率という.
メッセージの長さが k, 符号語の長さが n のとき, 符号化率 R は

$$R = \frac{k}{n}$$

である.
工学的な要請としては, 符号化率は1に近いほど良い.
※余分に送る（記録する）情報が少ないほど良いため.

問題：長さ3の単一パリティ検査符号 $\mathcal{C} = \{000, 101, 011, 110\}$ の符号化率はいくつ？

長さ3の単一パリティ検査符号の符号化率

符号化の説明のところで述べたように，長さ3の単一パリティ検査符号
$\mathcal{C} = \{000, 101, 011, 110\}$ の4つの符号語に割り当てられるビット列の長さ（メッセージ長）は
$k = 2$ であることから符号化率は　　である.

メッセージ	符号語
00	000
10	101
01	011
11	110

符号化は，符号語の重みが偶数になるように，メッセージの末尾に１ビット（パリティビットと呼ぶ）を付加することに対応.

余談：通信路容量と符号化率

シャノンは論文「通信の数学的理論」の中で，任意の反転誤り率が与えられたときに，誤りなしに通信可能な符号化率の上限を示した（通信路容量）.
しかし，シャノンの論文中に通信路容量を達成する具体的な符号は示されていない.
符号理論はシャノンの通信路容量を達成するために進化・発展してきた研究テーマである.
そして現在は，シャノンの通信路容量を達成する具体的な符号が複数知られている.

- Polar符号，Low-Density Parity-Check (LDPC)符号 など

理論的には通信路容量を達成する符号は示されているが，実用的にはそのような符号をできるだけ仮定を省いたうえで，現実的な計算量でいかに実現するかという問題が残されてる,

繰り返し符号(Repetition Code)

例：長さ3の繰り返し符号 $\mathcal{C} = \{000, 111\}$
長さ n の繰り返し符号は，長さ n の全て 0 もしくは 1 の2つの符号語を持つ.
つまり符号化は $0 \to 000, 1 \to 111$ となる.

単一パリティ検査符号は奇数の誤りを検出するのみで**訂正できなかった**が,
繰り返し符号は1ビットの誤りを検出し，**訂正することができる**.

問題：1ビットの誤りが発生し， 001 が読まれたとき．メッセージは 0 又は 1 のどちらか？

繰り返し符号(Repetition Code)

問題：1ビットの誤りが発生し，001 が読まれたとき．メッセージは 0 又は 1 のどちらか？
答え：
000 の一番右のビットを反転すると 001 になる
111 のどの1bitを反転しても 001 にならない

長さ3の繰り返し符号は1ビットの誤りを訂正可能
では，長さ5の繰り返し符号は何ビットの誤りを訂正可能でしょうか？

繰り返し符号の誤り訂正能力

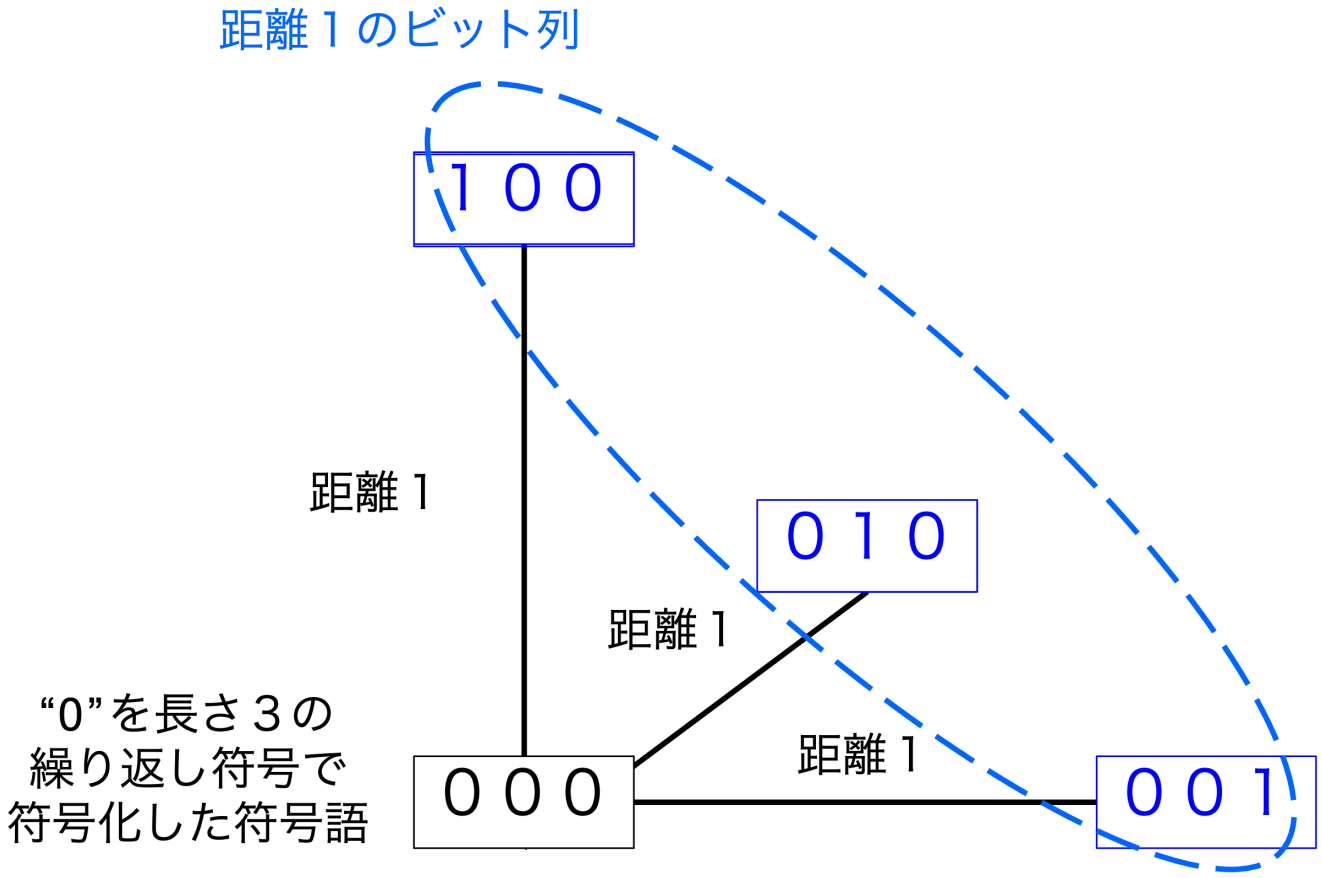
答え：

一般的には，長さ n の繰り返し符号は $\left\lfloor \frac{n-1}{2} \right\rfloor$ ビットの誤りを訂正できる．ただし $\lfloor x \rfloor$ は x 以下の最大の整数を表す．

誤り訂正能力の指標として重要な概念に，ハミング距離がある．次頁以降でハミング距離を説明する．

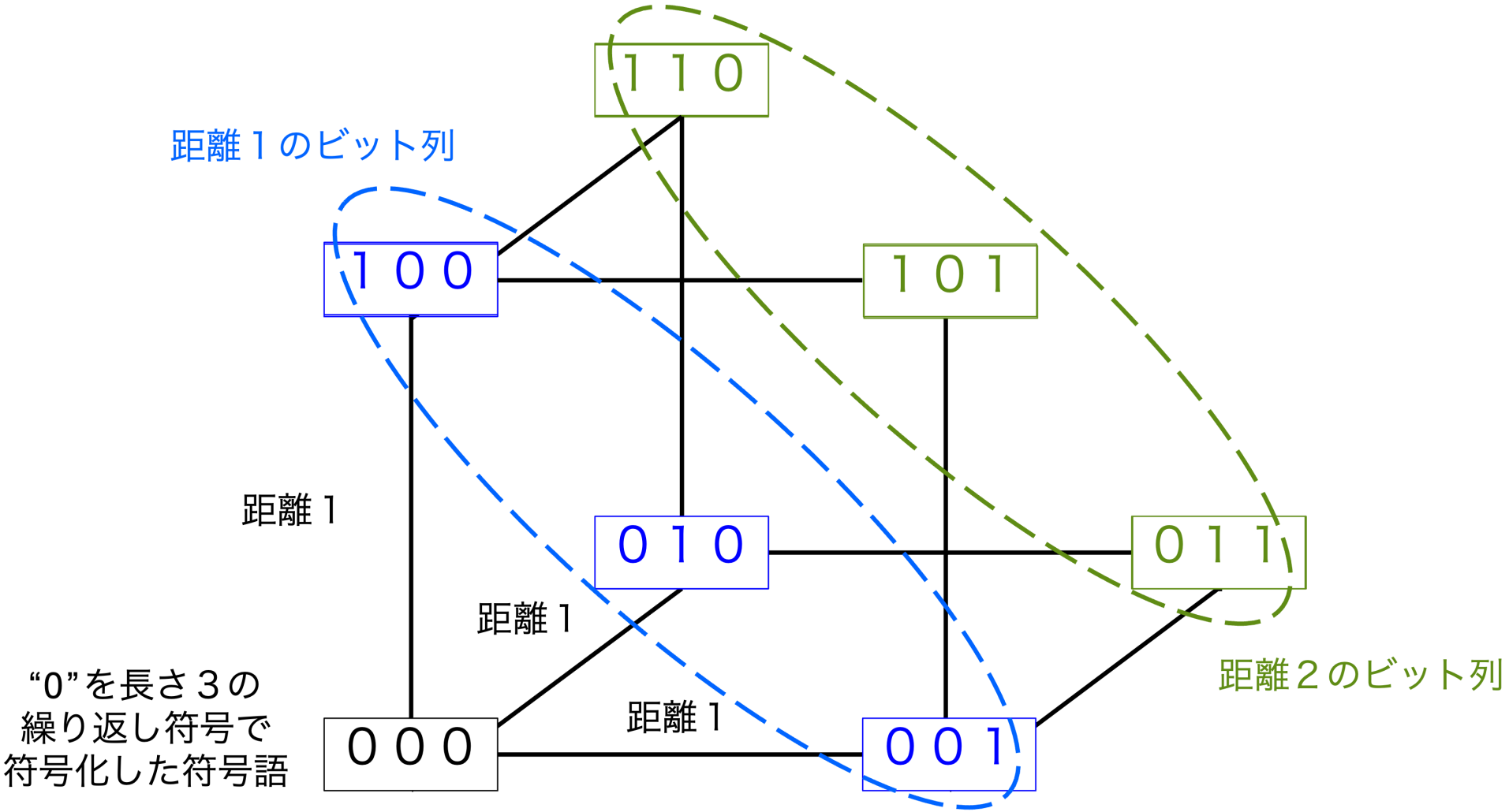
誤り訂正能力の指標：ハミング距離

ハミング距離とは，ビット列間の異なるビットの数である.



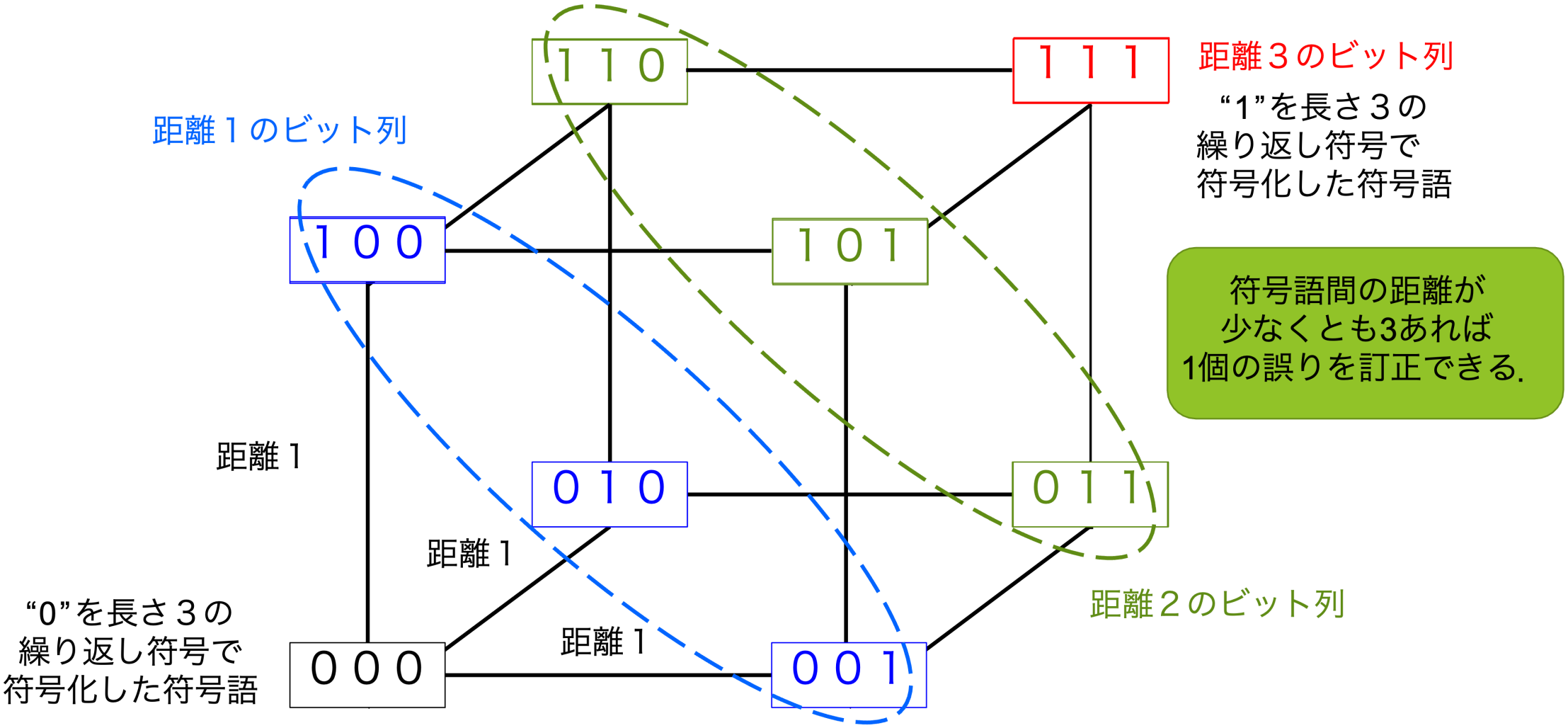
誤り訂正能力の指標：ハミング距離

ハミング距離とは，ビット列間の異なるビットの数である.



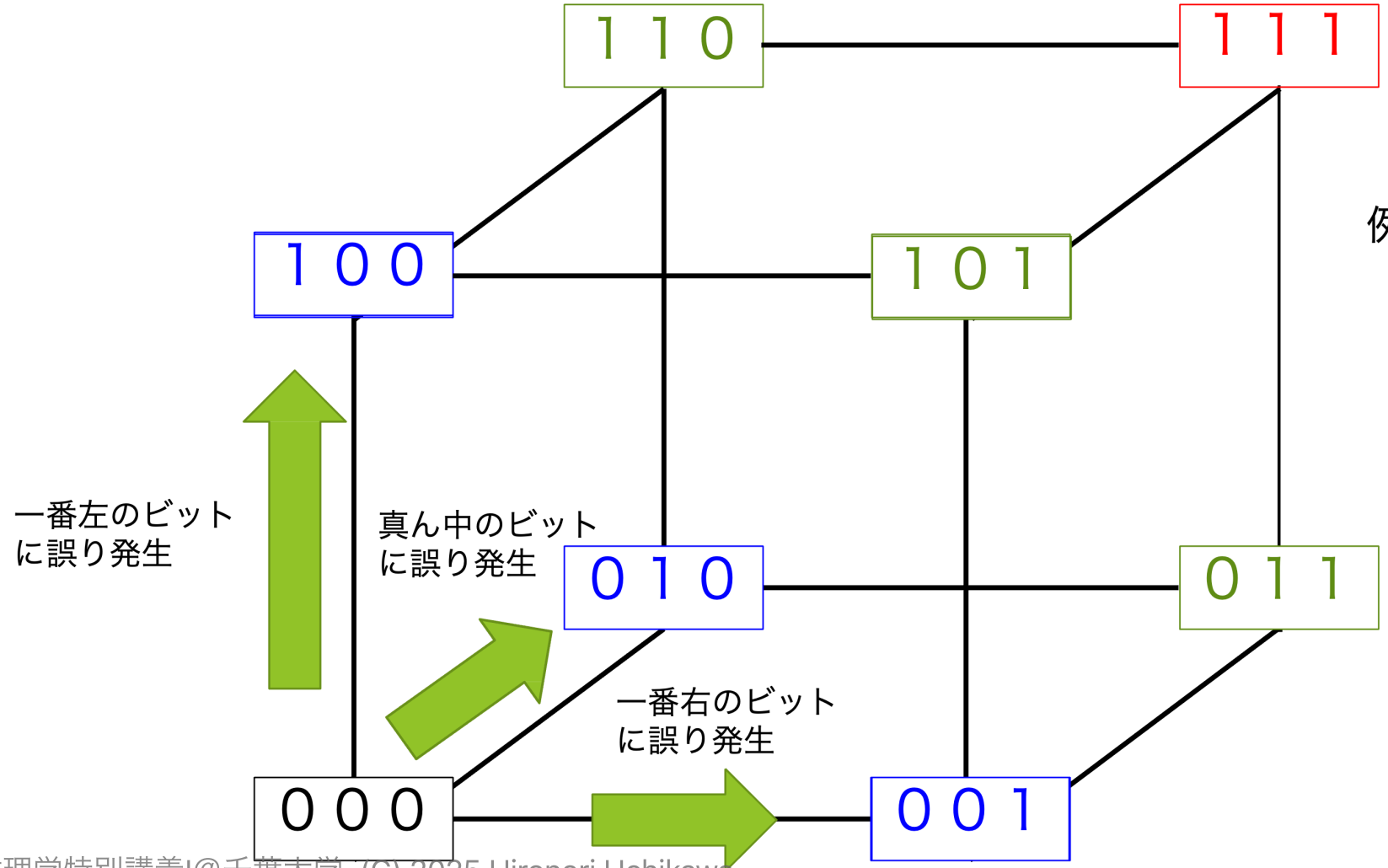
誤り訂正能力の指標：ハミング距離

ハミング距離とは，ビット列間の異なるビットの数である.



誤り訂正符号のコンセプト

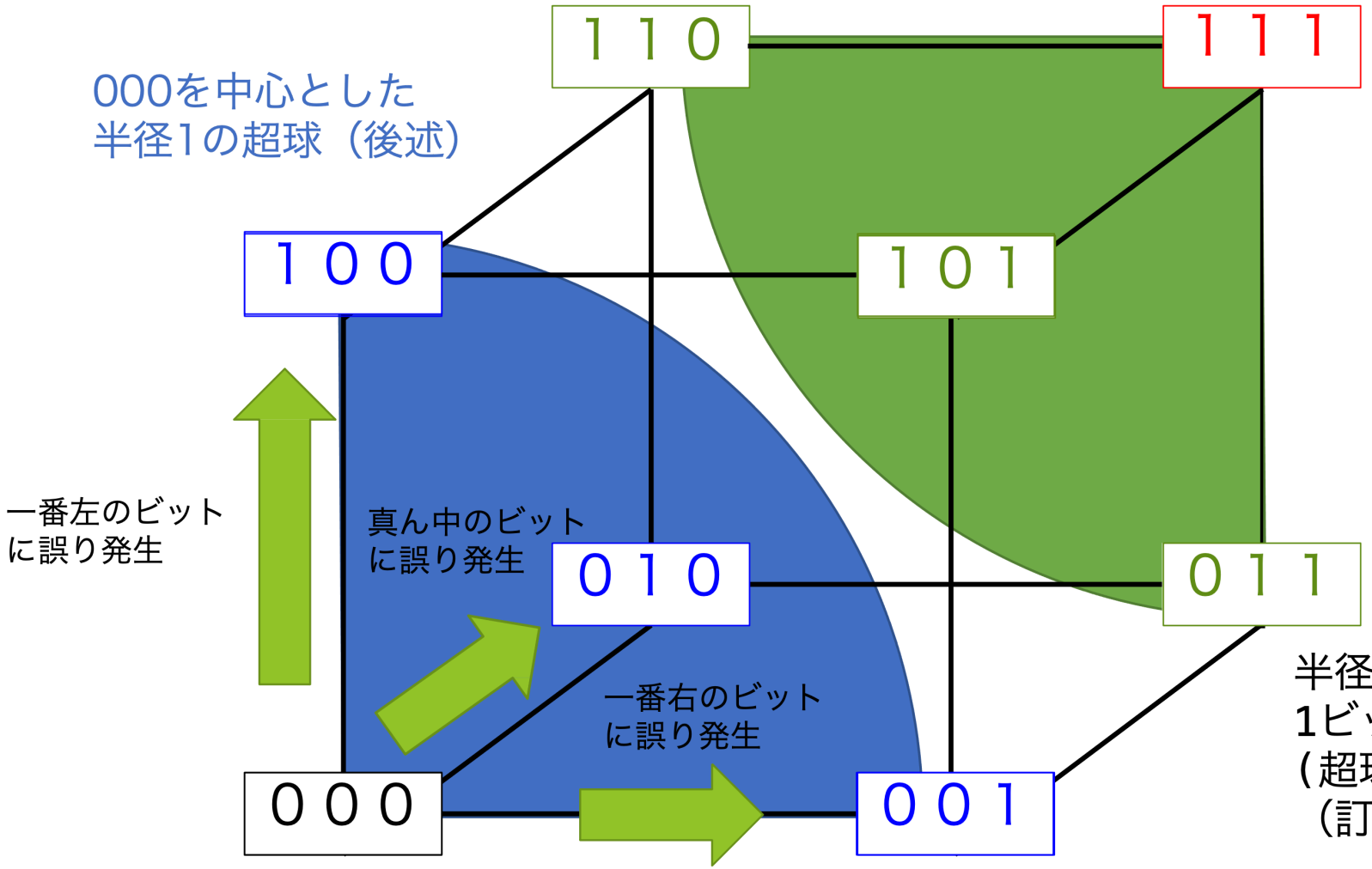
誤り訂正符号はメッセージ間のハミング距離を広げることで誤りへの耐性をもたせる.



例えば繰り返し符号では,
もともと距離が１だった
“0”と”1”を,
符号化によって
000と111とし,
距離を3に広げている.

誤り訂正符号のコンセプトと超球

誤り訂正符号はメッセージ間のハミング距離を広げることで誤りへの耐性をもたせる.



例えば繰り返し符号では,
もともと距離が１だった
“0”と”1”を,
符号化によって
000と111とし,
距離を3に広げている.

111を中心とした
半径1の超球

半径1の超球が重ならないため,
1ビットの誤りを訂正できる
(超球の中心の符号語に戻す
（訂正する）ことができる)

ハミング距離の定義

ある長さの n の2つのベクトル $\underline{a}$ と $\underline{b}$ があるとき，ベクトル $\underline{a}$ と $\underline{b}$ で要素が異なる座標の数をハミング距離と呼び，$d_H(\underline{a}, \underline{b})$と表記する.

ハミング距離は以下の距離の公理を満たす

1. 非負性：$d_H(\underline{a}, \underline{b}) \geq 0$
2. 一致性：$\underline{a} = \underline{b}$ ならば $d_H(\underline{a}, \underline{b}) = 0$
3. 対称性：$d_H(\underline{a}, \underline{b}) = d_H(\underline{b}, \underline{a})$
4. 三角不等式：$d_H(\underline{a}, \underline{c}) \leq d_H(\underline{a}, \underline{b}) + d_H(\underline{b}, \underline{c})$

符号の最小ハミング距離

長さ n の符号 $\mathcal{C}$ があるとき,

$$d = \min\left\{d_H(\underline{a}, \underline{b}) : \underline{a} \neq \underline{b},\ \underline{a}, \underline{b} \in \mathcal{C}\right\}$$

を符号 $\mathcal{C}$ の最小ハミング距離（以降はハミングを省略し最小距離）と呼ぶ

問題：長さ3の繰り返し符号 $\mathcal{C} = \{000, 111\}$ の最小距離は？
問題：長さ3の単一パリティ検査符号 $\mathcal{C} = \{000, 101, 011, 110\}$ の最小距離は？

符号 $\mathcal{C}$ の誤り訂正能力と最小距離の関係

命題

最小距離 d を持つ長さ n の符号 $\mathcal{C}$ は最大で $\left\lfloor \frac{d-1}{2} \right\rfloor$ 個の誤りを訂正できる．ただし $\lfloor x \rfloor$ は x を超えない最大の整数である．

証明に出てくる超球（Sphere）

ベクトル $\underline{a}$ から距離が高々 r のベクトルの集合を，$\underline{a}$ 中心とした半径 r の**超球**と呼ぶ．
証明では符号語 $\underline{a}$ を中心とした半径 $\left\lfloor \frac{d-1}{2} \right\rfloor$ の 2^k 個の超球を考える．k はメッセージ長．
符号語間の最小距離は d のため，これら全ての超球は重複しない．

$\frac{d-1}{2}$ までの誤りを訂正できることの証明

符号語 $\underline{a}$ が送信され，通信路で誤りが生じた（いくつかのビットがフリップした）受信ベクトル $\underline{r}$ が受信されたとする．さらに生じる誤りは $\left\lfloor \frac{d-1}{2} \right\rfloor$ 個を超えないと仮定する．

符号 $\mathcal{C}$ のそれぞれの符号語を中心とした半径 $\left\lfloor \frac{d-1}{2} \right\rfloor$ の 2^k 個の超球を考える．全ての超球は重複しないため，受信ベクトル $\underline{r}$ は唯一つの超球に含まれる．そしてその超球の中心は符号語 $\underline{a}$ である．
よって復号器は $\underline{r}$ が含まれる超球を探し，その中心のベクトルを復号ベクトルと出力することで，高々 $\left\lfloor \frac{d-1}{2} \right\rfloor$ 個の誤りは完全に訂正することができる．

$\frac{d-1}{2}$ を超える誤りは訂正できないことの証明

$\lfloor \frac{d-1}{2} \rfloor$ は符号が訂正できる最大の誤り数であることを証明する.

たとえば符号語間の距離 $d_H(\underline{a}, \underline{b}) = d$ となる符号語 $\underline{a}, \underline{b} \in \mathcal{C}$ を考えよう.
ここで, $d_H(\underline{a}, \underline{u}) = 1 + \lfloor \frac{d-1}{2} \rfloor$ かつ $d_H(\underline{b}, \underline{u}) = d - 1 - \lfloor \frac{d-1}{2} \rfloor$ となるベクトル $\underline{u}$ を用意する.
$d_H(\underline{b}, \underline{u}) \leq d_H(\underline{a}, \underline{u})$なので, もし $\underline{a}$ が送信されて $\underline{u}$ が受信されたとき
（つまり$1 + \lfloor \frac{d-1}{2} \rfloor$個の誤りが起きたとき),
復号器は正しく送信された符号語 $\underline{a}$ を復号することができない.
なぜならば符号語 $\underline{b}$ は少なくとも $\underline{a}$ と同じくらい $\underline{u}$ に近いからである.

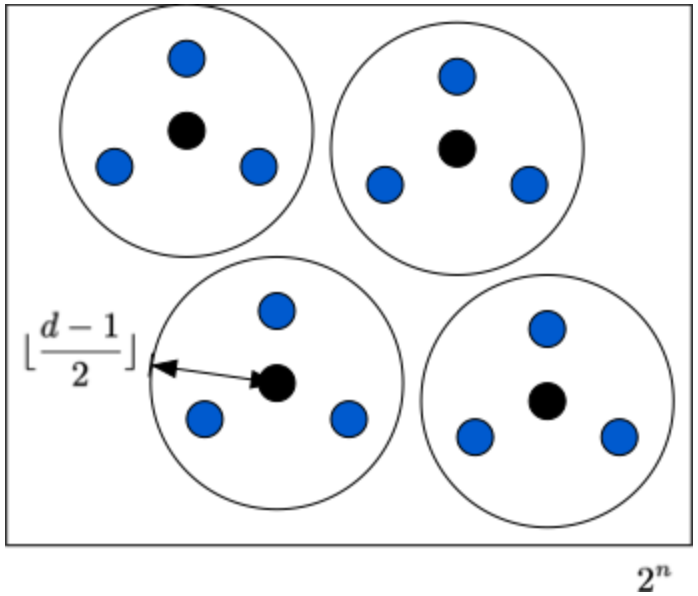
最小距離の限界式：ハミング限界

長さ n で最小距離 d の符号 $\mathcal{C}$ に含まれる符号語の数が 2^k (次元 k とも言う，メッセージの長さが k)のとき，

$$n - k \geq \log_2 V_n(\lfloor (d-1)/2 \rfloor)$$

が成り立つ．ただし，半径 r の超球のボリューム $V_n(r)$ は n 次元空間において，半径 r の超球に含まれるベクトルの数で $V_n(r) = \sum_{i=0}^{r} \binom{n}{i}$ である．

ハミング限界の証明



各符号語を中心とする 2^k 個の半径 $\lfloor (d-1)/2 \rfloor$ の超球は重複しないので，超球に含まれるベクトルの総数は $2^k V_n(\lfloor (d-1)/2 \rfloor)$ となり，その数は 2^n 以下であることから

$$2^n \geq 2^k V_n(\lfloor (d-1)/2 \rfloor).$$

ハミング限界の証明（続き）

$$2^n \geq 2^k V_n(\lfloor (d-1)/2 \rfloor).$$

底が 2 の対数をとり k を左辺に移項することで，ハミング限界

$$n - k \geq \log_2 V_n(\lfloor (d-1)/2 \rfloor)$$

が得られる.
ちなみに $n-k$ はパリティビットの長さ（冗長量）なので,
ハミング限界はある符号語の長さ n と最小距離 d が与えられたときのパリティビットの長さの下限を与えている.

ハミング限界の例

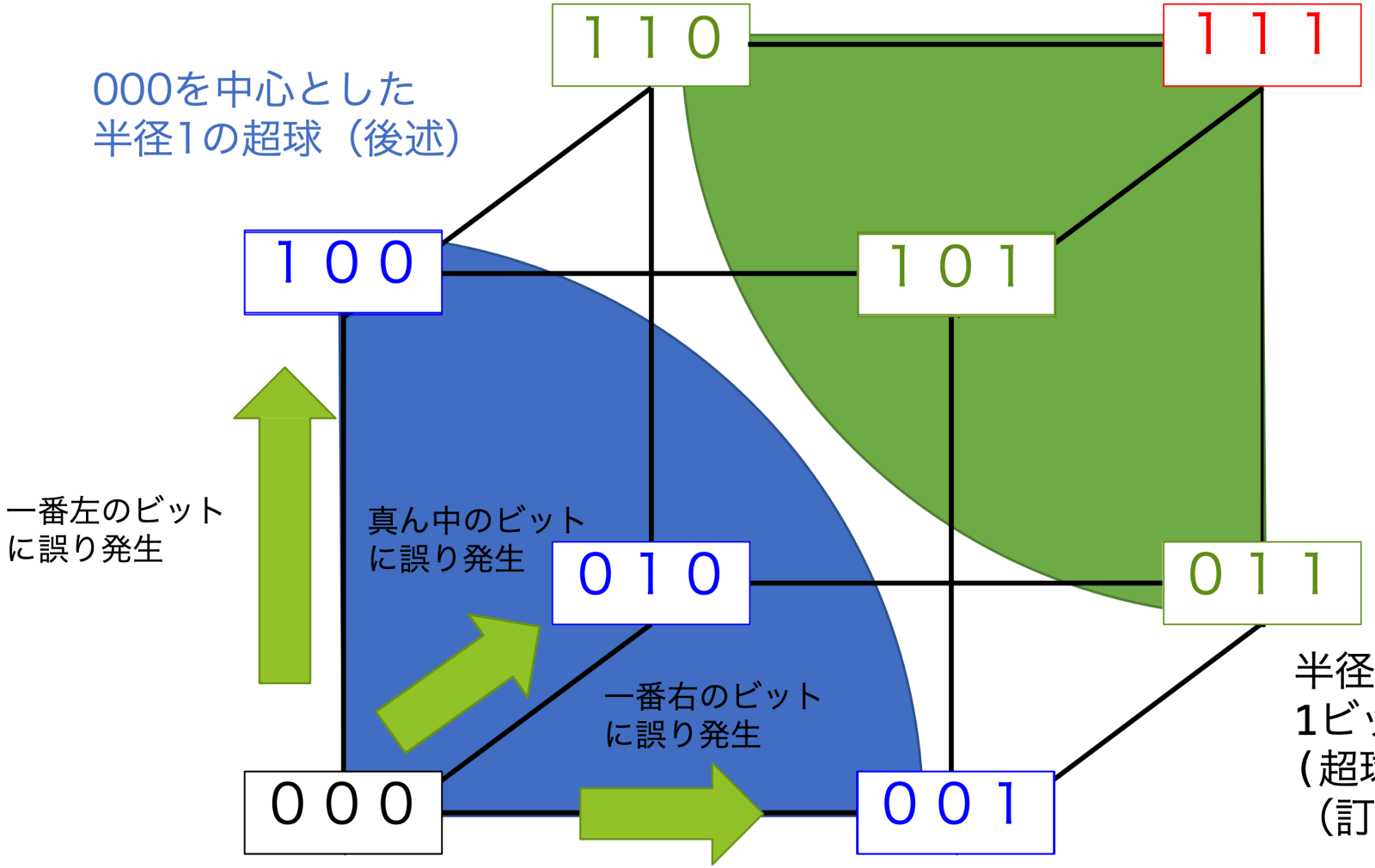
長さ 3 ($n=3$) の繰り返し符号 $\mathcal{C}=\{000,111\}$ は $k=1, d=3$.
ハミング限界 $n-k \geq \log_2 V_n(\lfloor (d-1)/2 \rfloor)$ の
左辺は $3-1=2$.
右辺の超球のボリュームが

$$V_3(\lfloor (3-1)/2 \rfloor) = V_3(1) = \sum_{i=0}^{1} \binom{3}{i} = 1 + \binom{3}{1} = 1+3 = 4$$

なので，$\log_2 4 = 2$となり，左辺と右辺が一致する.
これは先の図で超級が3次元空間を埋め尽くしていたことからもわかる.
一般にハミング限界を満たす符号（n次元空間を半径$\lfloor (d-1)/2 \rfloor$ の超球で埋め尽くす）を**完全符号**と呼ぶ.

誤り訂正符号のコンセプトと超球（再掲）

誤り訂正符号はメッセージ間のハミング距離を広げることで誤りへの耐性をもたせる.



例えば繰り返し符号では,
もともと距離が１だった
“0”と”1”を,
符号化によって
000と111とし,
距離を3に広げている.

111を中心とした
半径1の超球

半径1の超球が重ならないため,
1ビットの誤りを訂正できる
(超球の中心の符号語に戻す
（訂正する）ことができる)

まとめ

- 符号とは符号語(codeword)と呼ばれるビット列の集合
- 単一パリティ検査符号は符号語に含まれる1の数（**重み**と言う）が偶数の符号語の集合，**奇数ビットの誤りが検出できる.**
- 符号化率 R は符号語に占めるメッセージの割合 $\frac{k}{n}$ である.
- 誤り訂正符号はメッセージ間のハミング距離を広げることで誤りへの耐性をもたせる.
- 符号の最小ハミング距離（最小距離）と訂正能力の関係.
- ハミング限界
  - 符号語の長さ n と最小距離 d が与えられたときのパリティビットの長さの下限.
  - 符号の長さ n とメッセージ長 k が与えられたときに達成できる最小距離の上限（訂正能力の限界）.